```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>

<body>

<p>Hey, diddle, diddle, the cat and the fiddle.</p>
<p>Starlight, starbright, first star I see tonight,<br>
 I wish I may, I wish I might, have the wish I wish tonight<br>
</p>
<p>The quick brown fox slyly jumped over the lazy dog.</p>

</body>
</html>
```

Figure 1

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>

<body>
<p>Mary had a little lamb, its fleece was white as snow.</p>
<p>Hey, diddle, diddle, the cat and the fiddle.</p>
<p>Starlight, starbright, first star I see tonight,<br>
 I wish I may, I wish I might, have the wish I wish tonight<br>
</p>
<p>The quick brown fox slyly jumped over the lazy dog.</p>

</body>
</html>
```

Figure 2

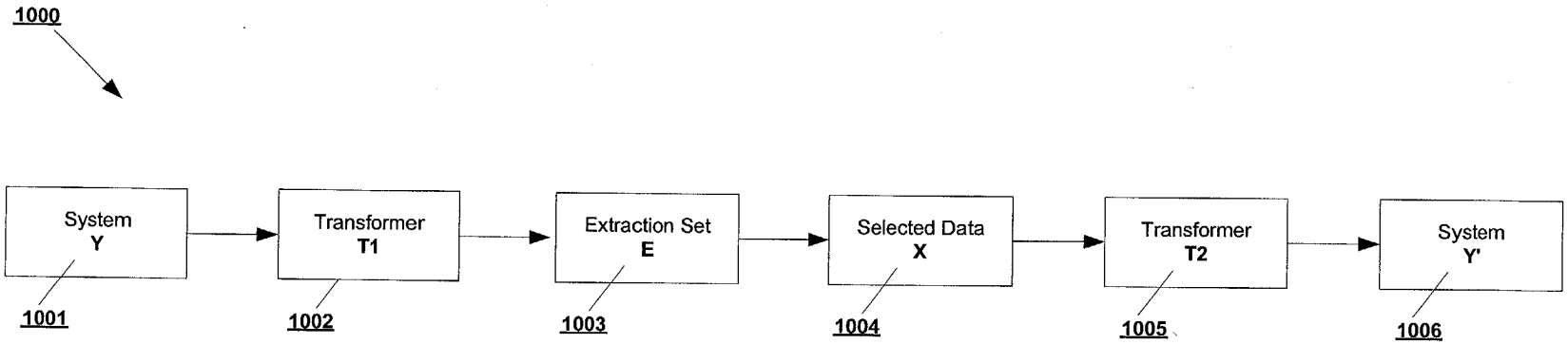
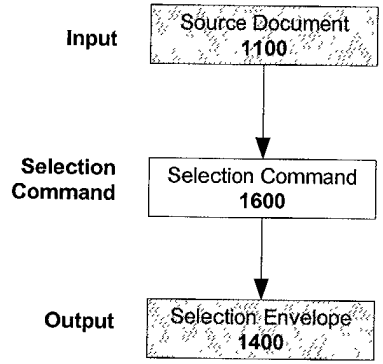

1000
System
Y
1001
Transformer
T1
1002
Extraction Set
E
1003
Selected Data
X
1004
Transformer
T2
1005
System
Y'
1006

Figure 3


Input
Source Document
1100
Selection
Command
Selection Command
1600
Output
Selection Envelope
1400

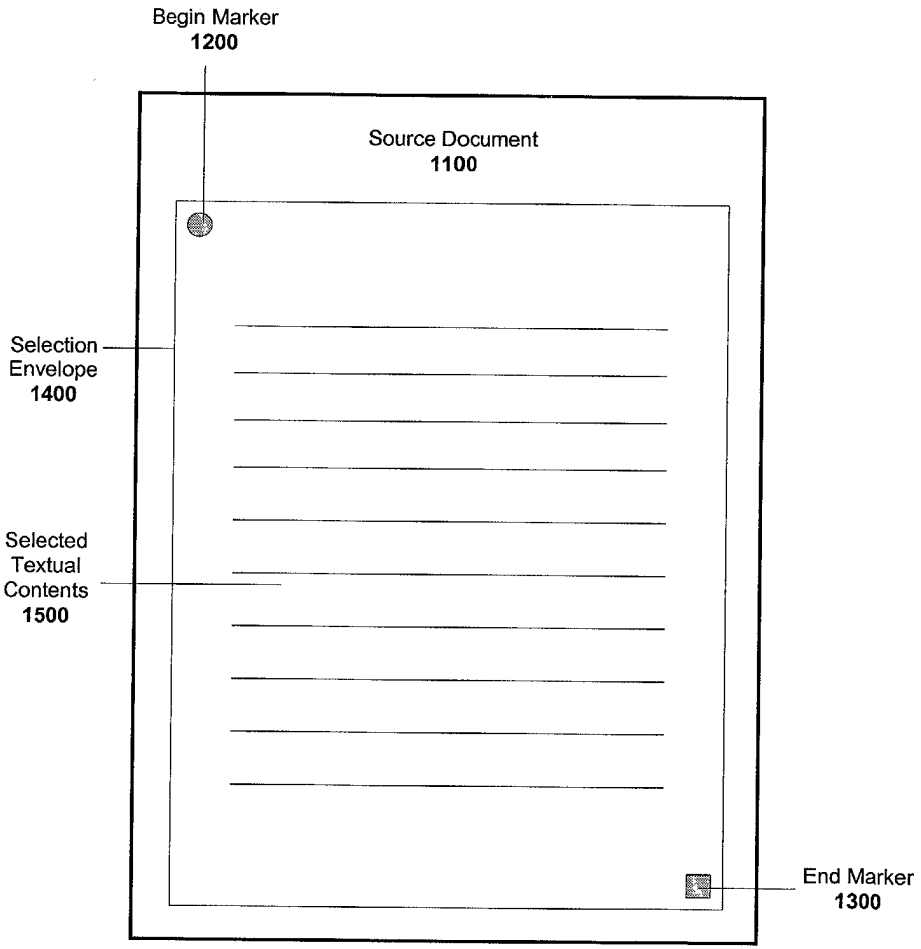

Begin Marker
1200
Source Document
1100
Selection
Envelope
1400
Selected
Textual
Contents
1500
End Marker
1300

Figure 4

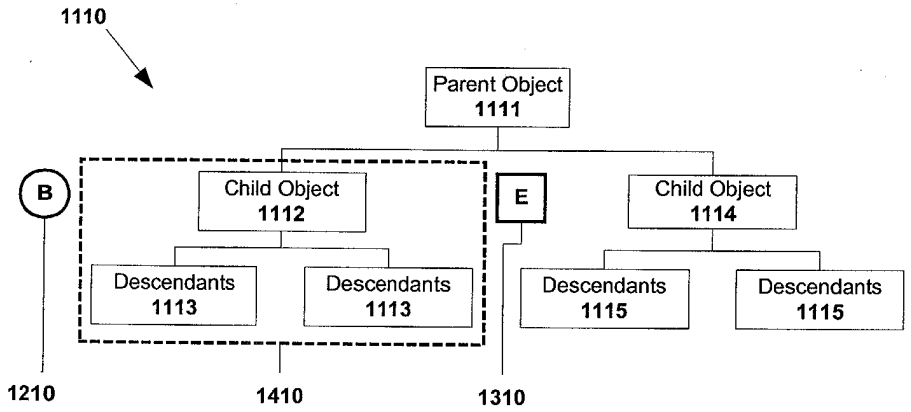


Figure 5

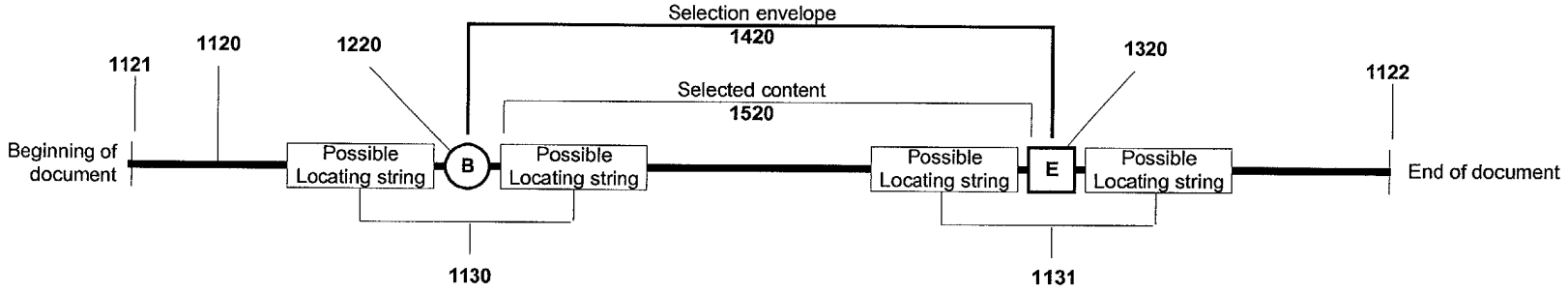
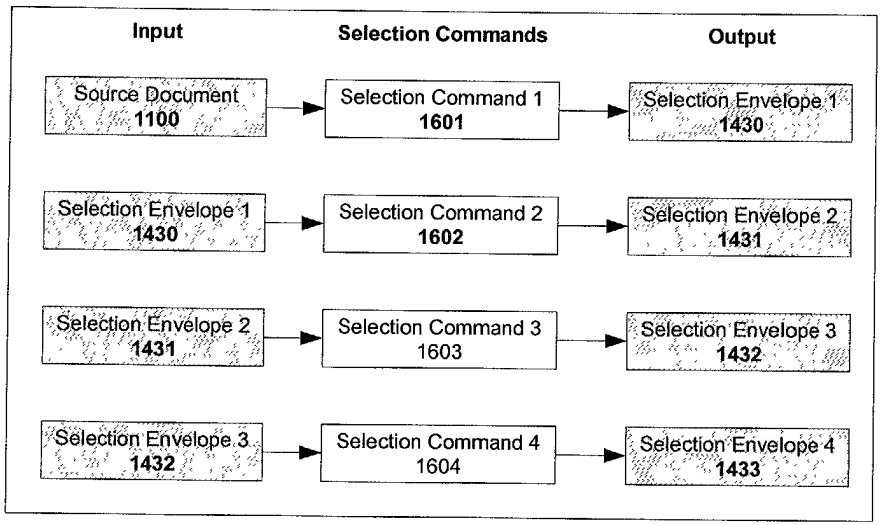


Figure 6


Input
Selection Commands
Output
Source Document 1100
Selection Command 1 1601
Selection Envelope 1 1430
Selection Envelope 1 1430
Selection Command 2 1602
Selection Envelope 2 1431
Selection Envelope 2 1431
Selection Command 3 1603
Selection Envelope 3 1432
Selection Envelope 3 1432
Selection Command 4 1604
Selection Envelope 4 1433

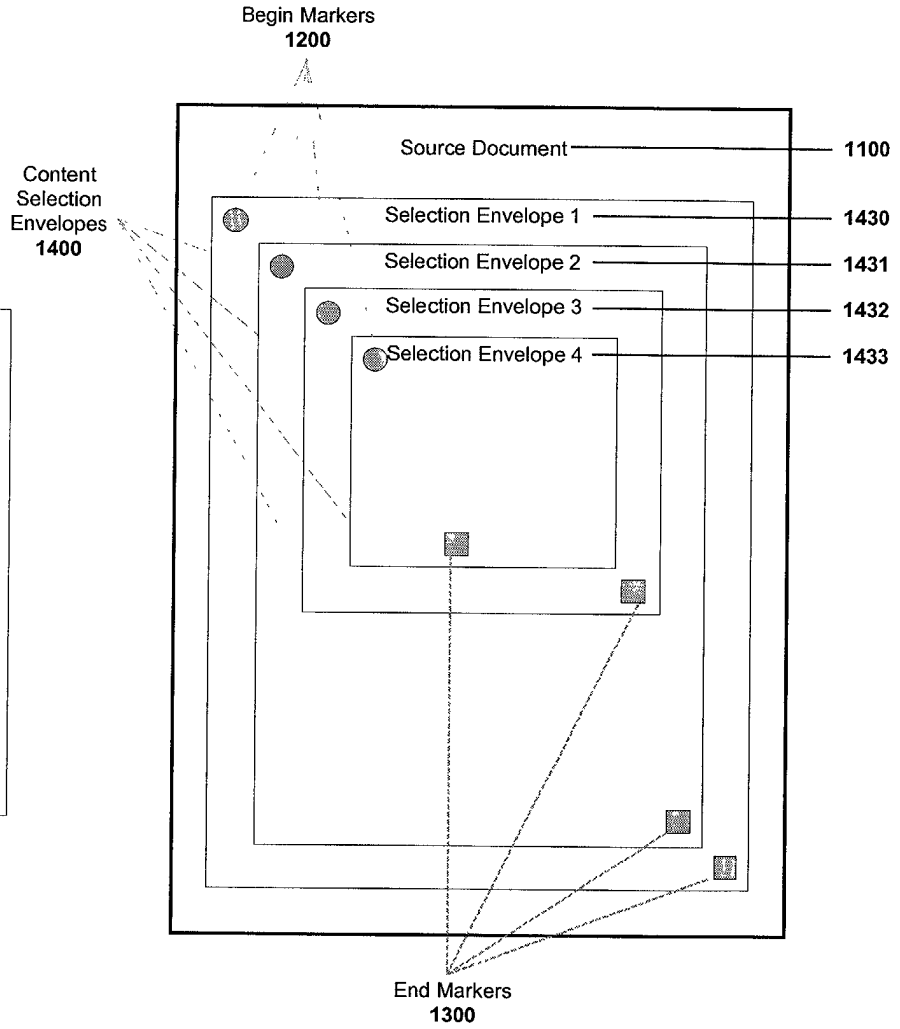

Begin Markers
1200
Content Selection Envelopes
1400
Source Document
1100
Selection Envelope 1
1430
Selection Envelope 2
1431
Selection Envelope 3
1432
Selection Envelope 4
1433
End Markers
1300

Figure 7

20230 DOCSOOT

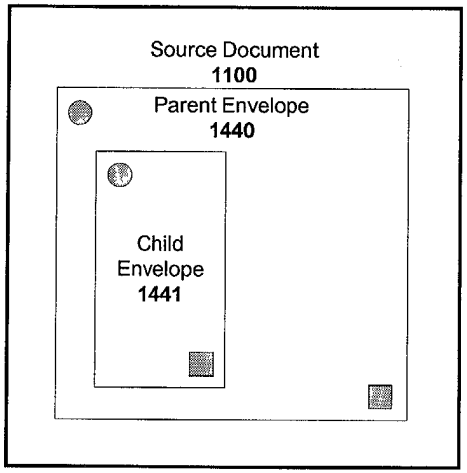


Figure 8

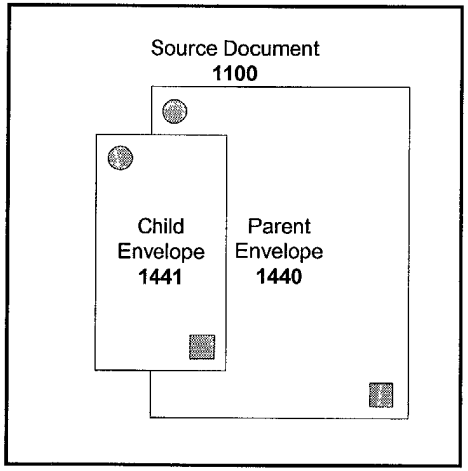


Figure 9

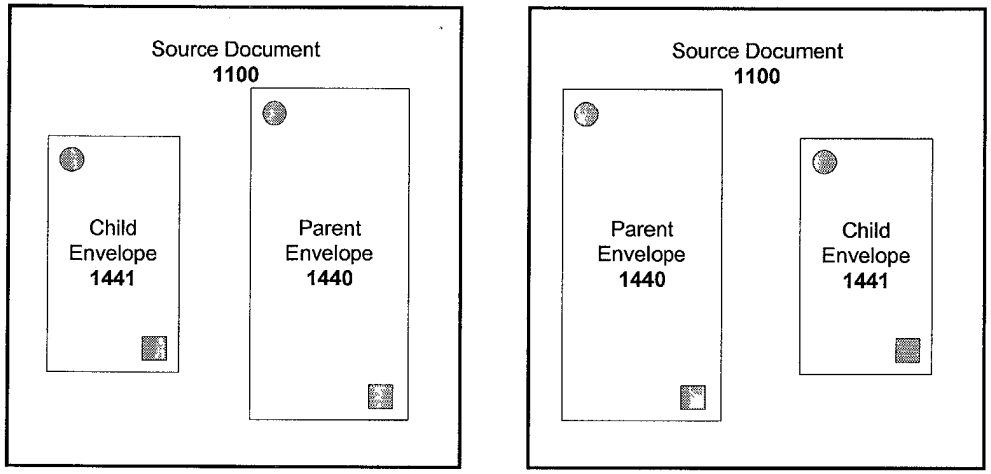


Figure 10

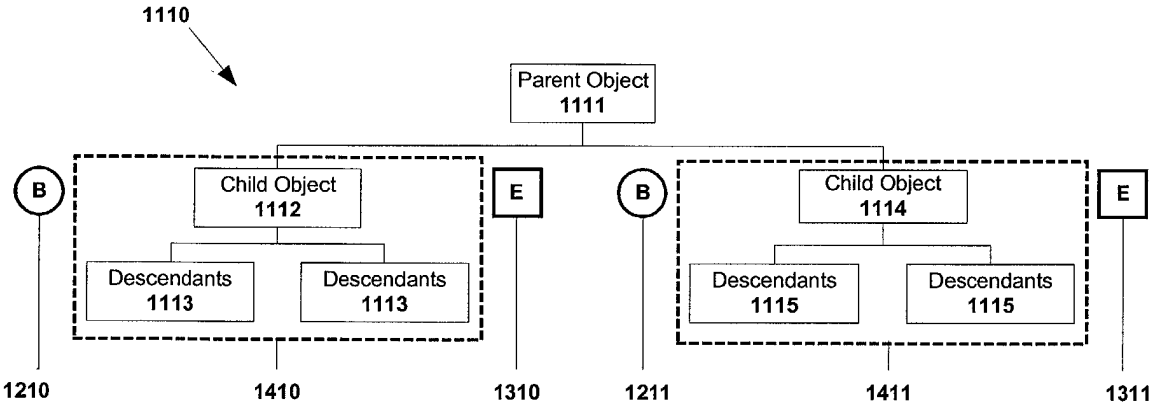


Figure 11

20220706 085900T

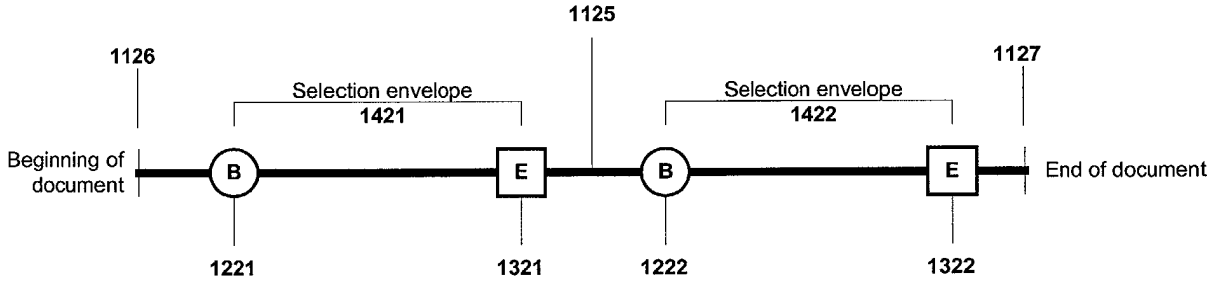


Figure 12

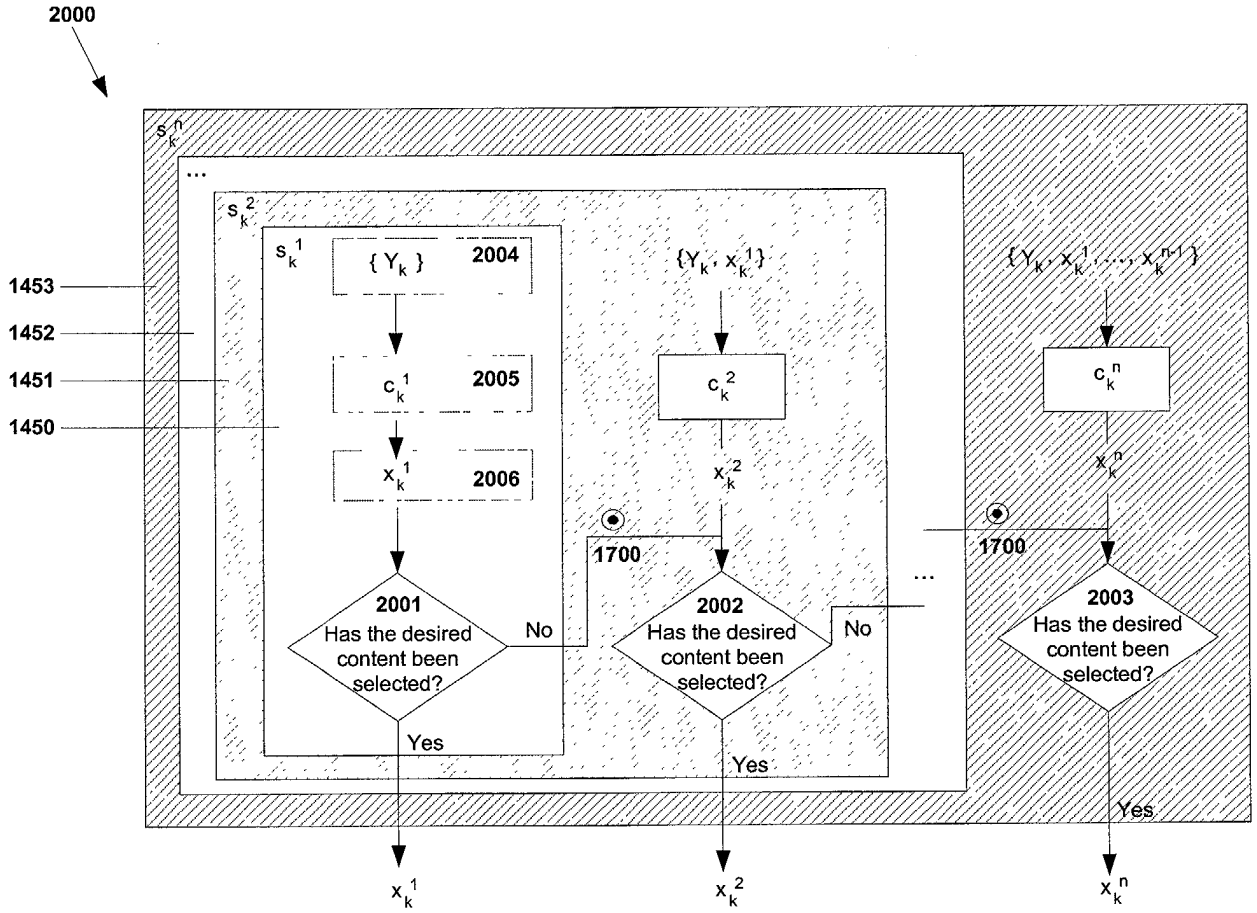

2000
1453
1452
1451
1450
s_k^n
s_k^2
s_k^1
{ Y_k }
2004
c_k^1
2005
x_k^1
2006
2001
Has the desired content been selected?
No
Yes
x_k^1
1700
$\{Y_k, x_k^1\}$
c_k^2
x_k^2
2002
Has the desired content been selected?
No
Yes
x_k^2
...
1700
$\{Y_k, x_k^1, ..., x_k^{n-1}\}$
c_k^n
x_k^n
2003
Has the desired content been selected?
Yes
x_k^n

Figure 13A

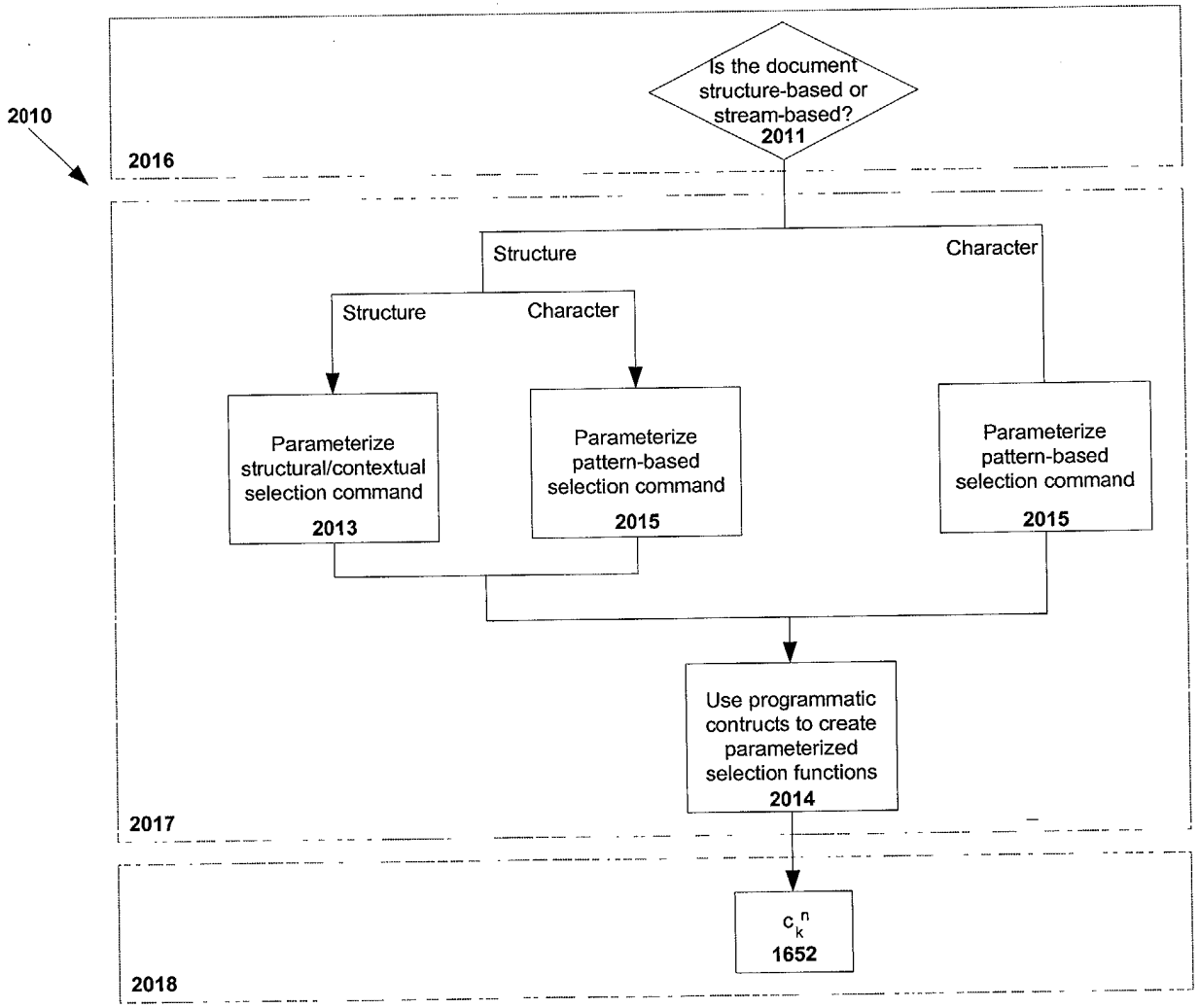

2010
2016
Is the document structure-based or stream-based?
2011
Structure
Character
Structure
Character
Parameterize structural/contextual selection command
2013
Parameterize pattern-based selection command
2015
Parameterize pattern-based selection command
2015
Use programmatic contructs to create parameterized selection functions
2014
2017
c_k^n
1652
2018

Figure 13B

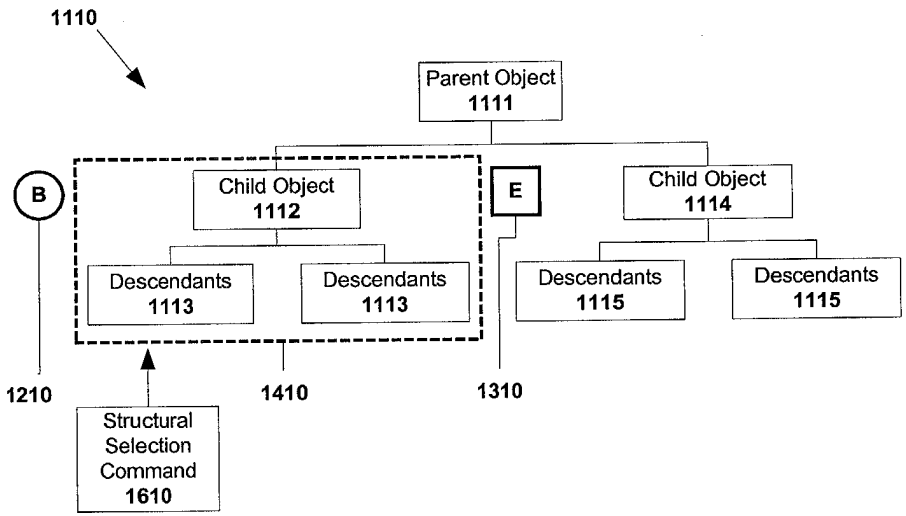


Figure 14

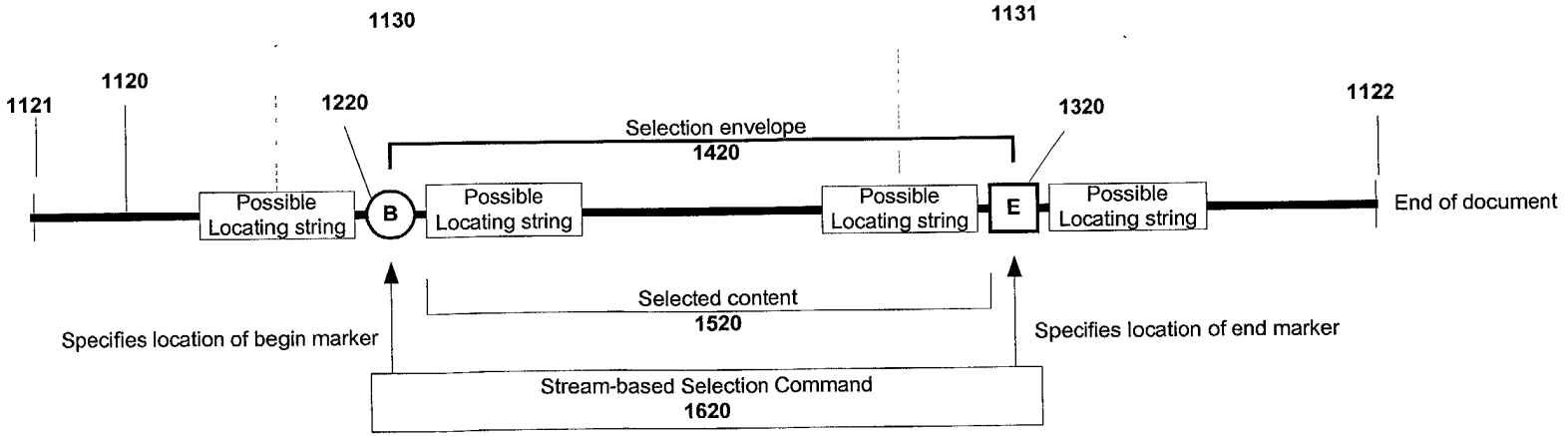


Figure 15

20220T0" 0ES90QT

Home News Product Contact Search

Page Title

This is a page to demonstrate the INE approach for content identification and extraction

Section Title

The INE approach for content identification and extraction is quite robust due to its multi-step, nested content selection system This system narrows each step's scope of selection relative to the last and also allows the selection to be defined in one of multiple dimensions Content selection are created by defining a start and end marker within a body of HTML These markers define an envelope of content

Column 1	Column 2	Column 3	Column 4
Content	Content	Content	Content
Content	Content	Content	Row1 Row2

Figure 16

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html>
<head>
        <title>Sample Page</title>
</head>

<body>
<P><STRONG>
<TABLE>

  <TR>
    <TD>Home</TD>
    <TD>News</TD>
    <TD>Product</TD>
    <TD>Contact</TD>
    <TD>Search</TD></TR></TABLE>  </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and
extraction.</P>
<P><STRONG>Section Title</STRONG></P>
<P>The INE approach for content identification and extraction is quite robust
due to its multi-step, nested content ,selection system  This
system narrows each step's scope of selection relative to the last and also
allows the selection to be defined in one of multiple
dimensions  Content selection are created by defining a start and end
marker within a body of HTML. These markers define an envelope of content. </P>
 <TABLE>
  <TR>
    <TD>Column 1</TD>
    <TD>Column 2</TD>
    <TD>Column 3</TD>
    <TD>Column 4</TD></TR>
  <TR>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD></TR>
  <TR>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>
      <TABLE>
        <TR>
          <TD>Row1</TD></TR>
        <TR>
          <TD>Row2</TD></TR></TABLE></TD></TR>
</TABLE>
<P> ,</P>
<P> ,</P>

</body>
</html>
```

Figure 17

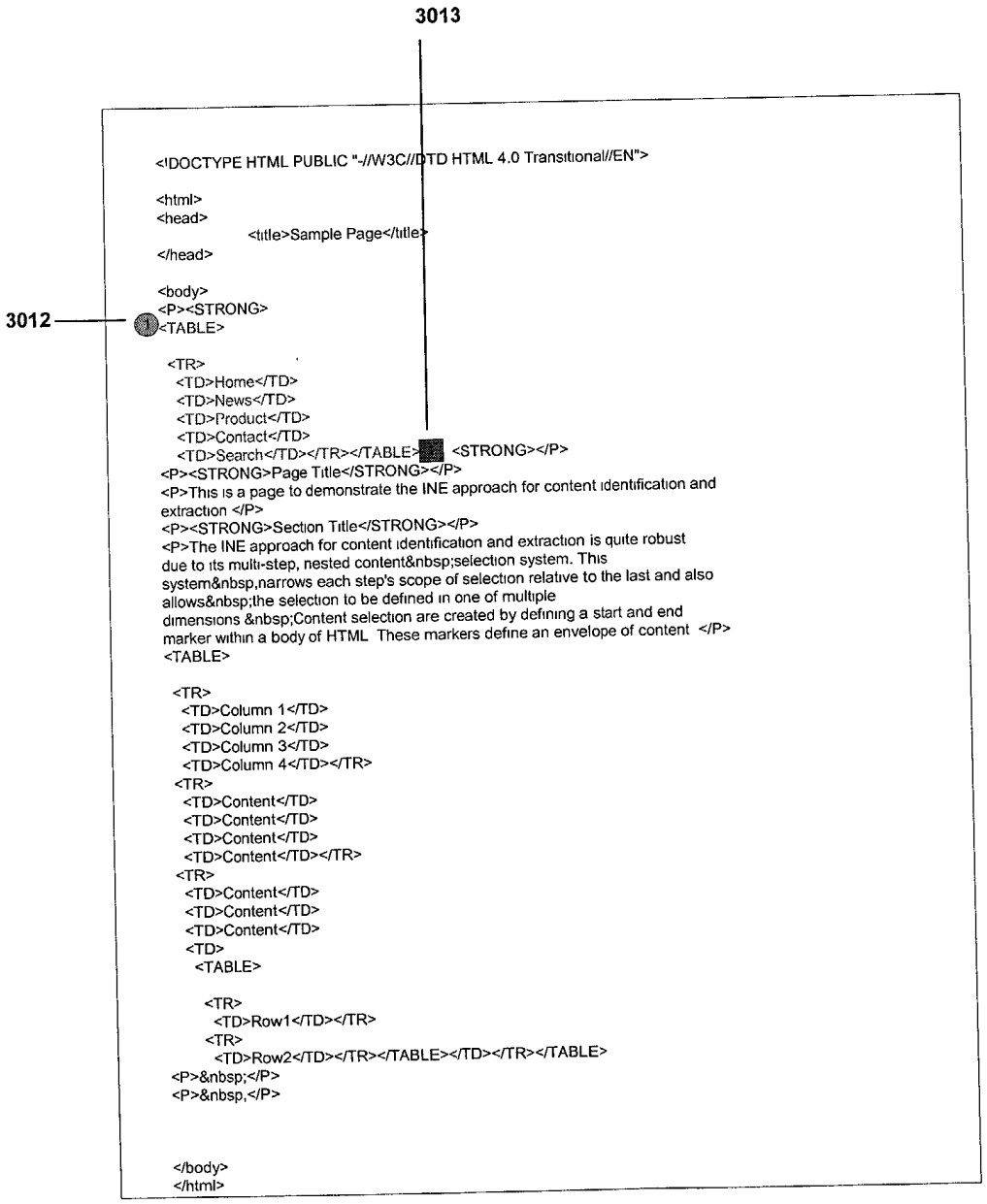
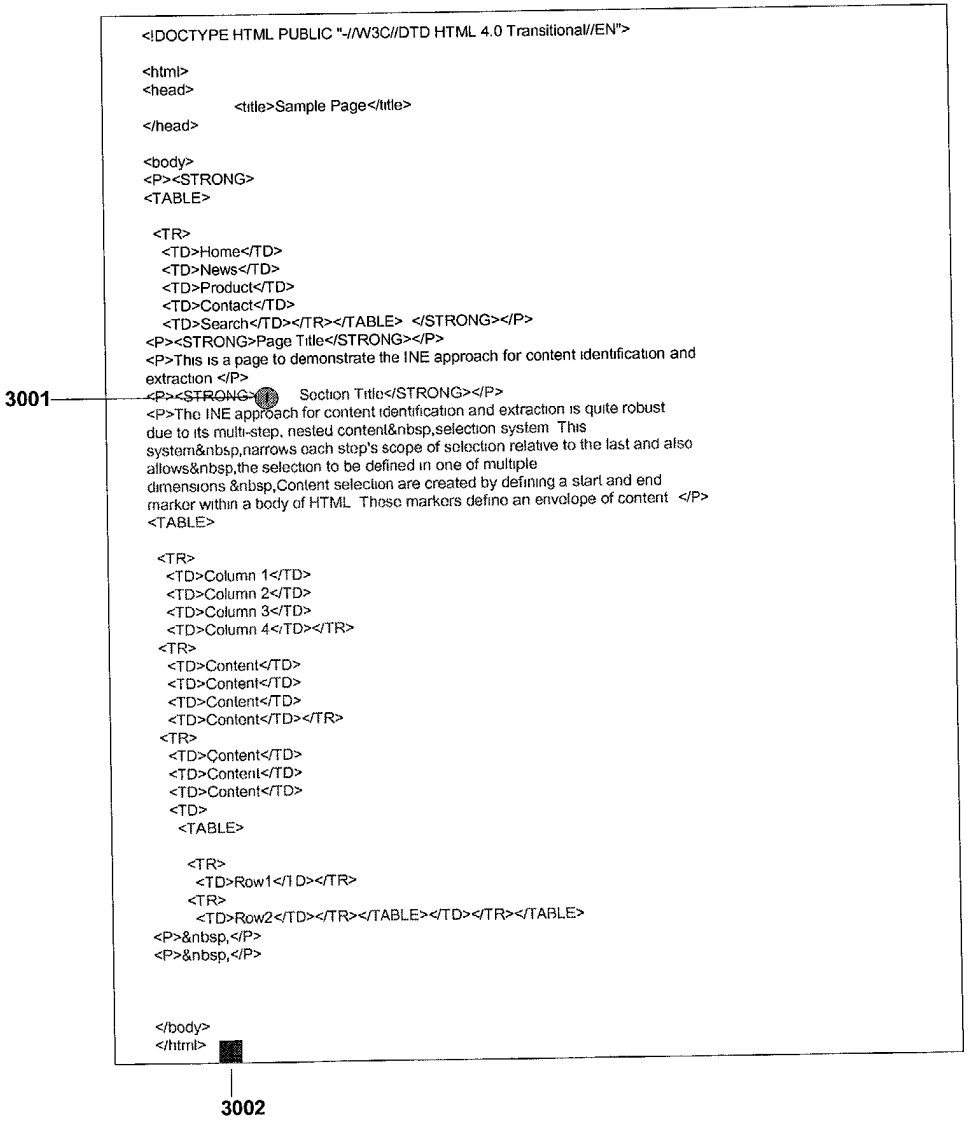

3013
3012
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<title>Sample Page</title>
</head>
<body>
<P><STRONG>
<TABLE>
<TR>
<TD>Home</TD>
<TD>News</TD>
<TD>Product</TD>
<TD>Contact</TD>
<TD>Search</TD></TR></TABLE> <STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and extraction </P>
<P><STRONG>Section Title</STRONG></P>
<P>The INE approach for content identification and extraction is quite robust due to its multi-step, nested content selection system. This system ,narrows each step's scope of selection relative to the last and also allows the selection to be defined in one of multiple dimensions Content selection are created by defining a start and end marker within a body of HTML These markers define an envelope of content </P>
<TABLE>
<TR>
<TD>Column 1</TD>
<TD>Column 2</TD>
<TD>Column 3</TD>
<TD>Column 4</TD></TR>
<TR>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD></TR>
<TR>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD>
<TD>
<TABLE>
<TR>
<TD>Row1</TD></TR>
<TR>
<TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> </P>
<P> ,</P>
</body>
</html>

Figure 18

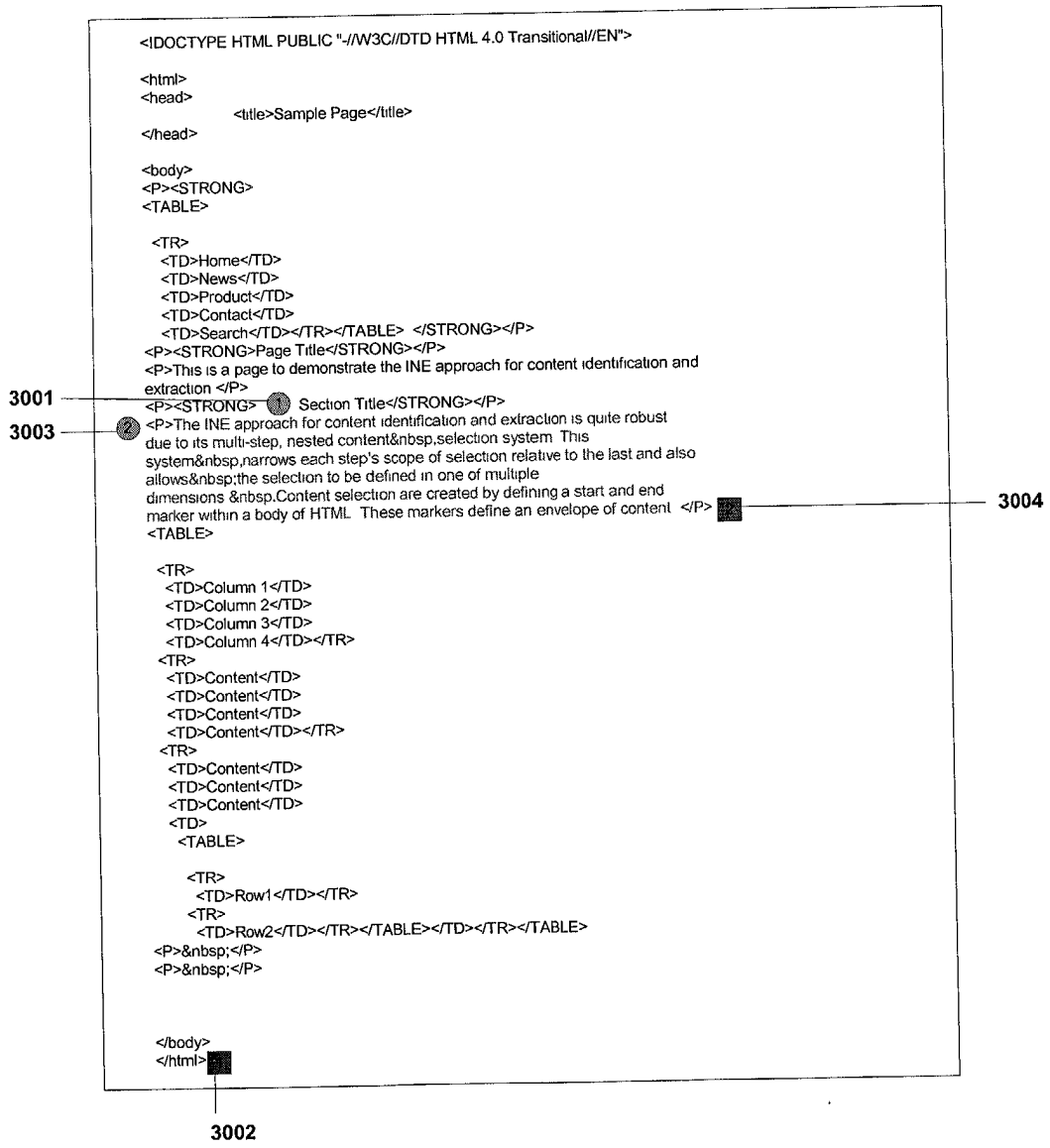
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html>
<head>
        <title>Sample Page</title>
</head>

<body>
<P><STRONG>
<TABLE>

  <TR>
    <TD>Home</TD>
    <TD>News</TD>
    <TD>Product</TD>
    <TD>Contact</TD>
    <TD>Search</TD></TR></TABLE>  </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and
extraction </P>
<P><STRONG>Section Title</STRONG></P>
<P>The INE approach for content identification and extraction is quite robust
due to its multi-step, nested content selection system  This
system narrows each step's scope of selection relative to the last and also
allows the selection to be defined in one of multiple
dimensions. Content selection are created by defining a start and end
marker within a body of HTML. These markers define an envelope of content. </P>
<TABLE>

  <TR>
    <TD>Column 1</TD>
    <TD>Column 2</TD>
    <TD>Column 3</TD>
    <TD>Column 4</TD></TR>
  <TR>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD></TR>
  <TR>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>Content</TD>
    <TD>
      <TABLE>

        <TR>
          <TD>Row1</TD></TR>
        <TR>
          <TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> </P>
<P> </P>



</body>
</html>
```

Figure 19




```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html>
<head>
        <title>Sample Page</title>
</head>

<body>
<P><STRONG>
<TABLE>

 <TR>
  <TD>Home</TD>
  <TD>News</TD>
  <TD>Product</TD>
  <TD>Contact</TD>
  <TD>Search</TD></TR></TABLE>  </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and
extraction </P>
<P><STRONG>1    Section Title</STRONG></P>
<P>The INE approach for content identification and extraction is quite robust
due to its multi-step, nested content ,selection system  This
system ,narrows each step's scope of selection relative to the last and also
allows ,the selection to be defined in one of multiple
dimensions  ,Content selection are created by defining a start and end
marker within a body of HTML  These markers define an envelope of content  </P>
<TABLE>

 <TR>
  <TD>Column 1</TD>
  <TD>Column 2</TD>
  <TD>Column 3</TD>
  <TD>Column 4</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>
   <TABLE>

    <TR>
     <TD>Row1</TD></TR>
    <TR>
     <TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> ,</P>
<P> ,</P>


</body>
</html>
```

Figure 20




```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html>
<head>
        <title>Sample Page</title>
</head>

<body>
<P><STRONG>
<TABLE>

 <TR>
  <TD>Home</TD>
  <TD>News</TD>
  <TD>Product</TD>
  <TD>Contact</TD>
  <TD>Search</TD></TR></TABLE> </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and
extraction </P>
<P><STRONG> 1 Section Title</STRONG></P>
2 <P>The INE approach for content identification and extraction is quite robust
due to its multi-step, nested content selection system  This
system narrows each step's scope of selection relative to the last and also
allows the selection to be defined in one of multiple
dimensions  .Content selection are created by defining a start and end
marker within a body of HTML  These markers define an envelope of content  </P>
<TABLE>

 <TR>
  <TD>Column 1</TD>
  <TD>Column 2</TD>
  <TD>Column 3</TD>
  <TD>Column 4</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>
   <TABLE>

    <TR>
     <TD>Row1</TD></TR>
    <TR>
     <TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> </P>
<P> </P>


</body>
</html>
```

Figure 21

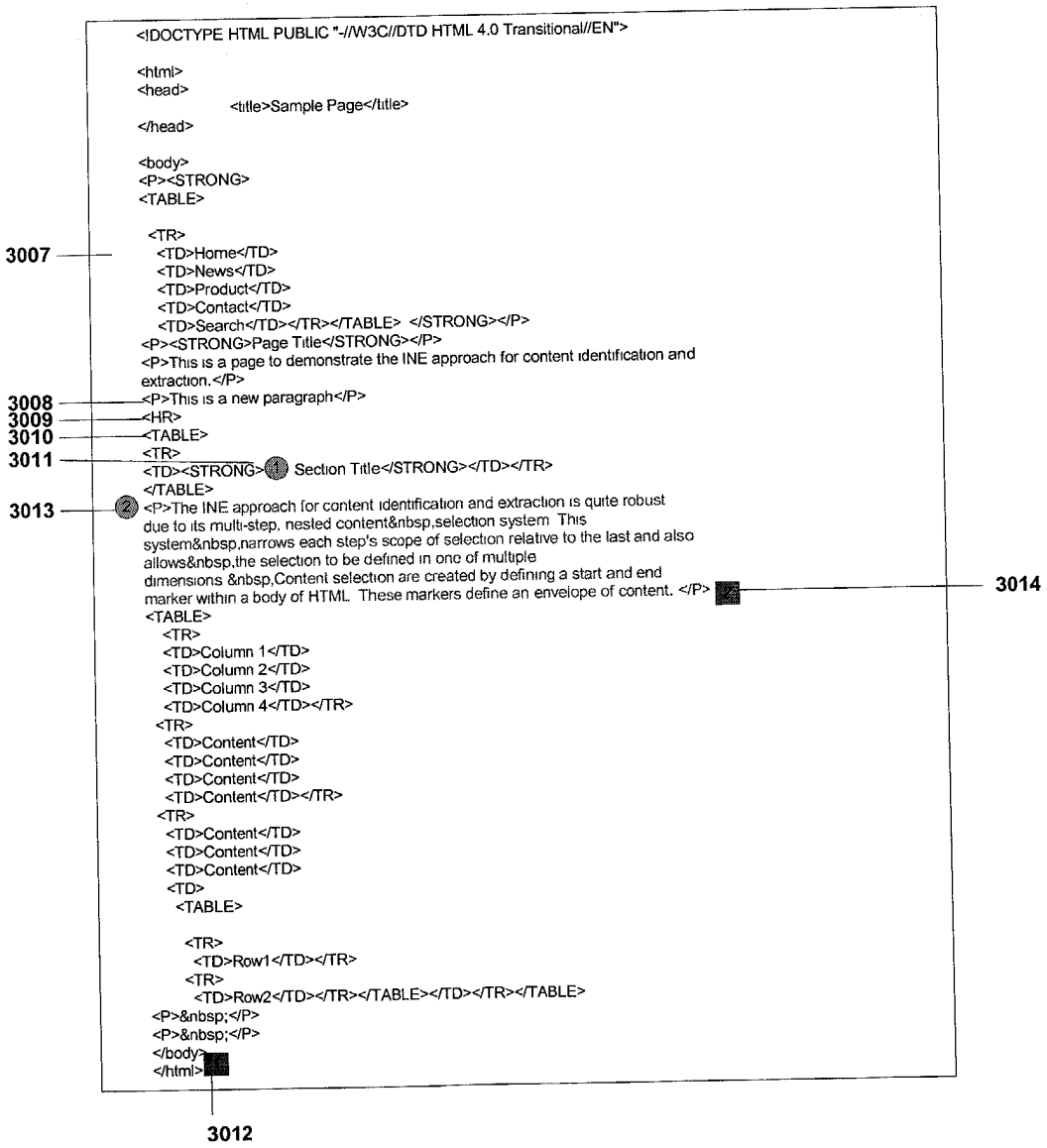
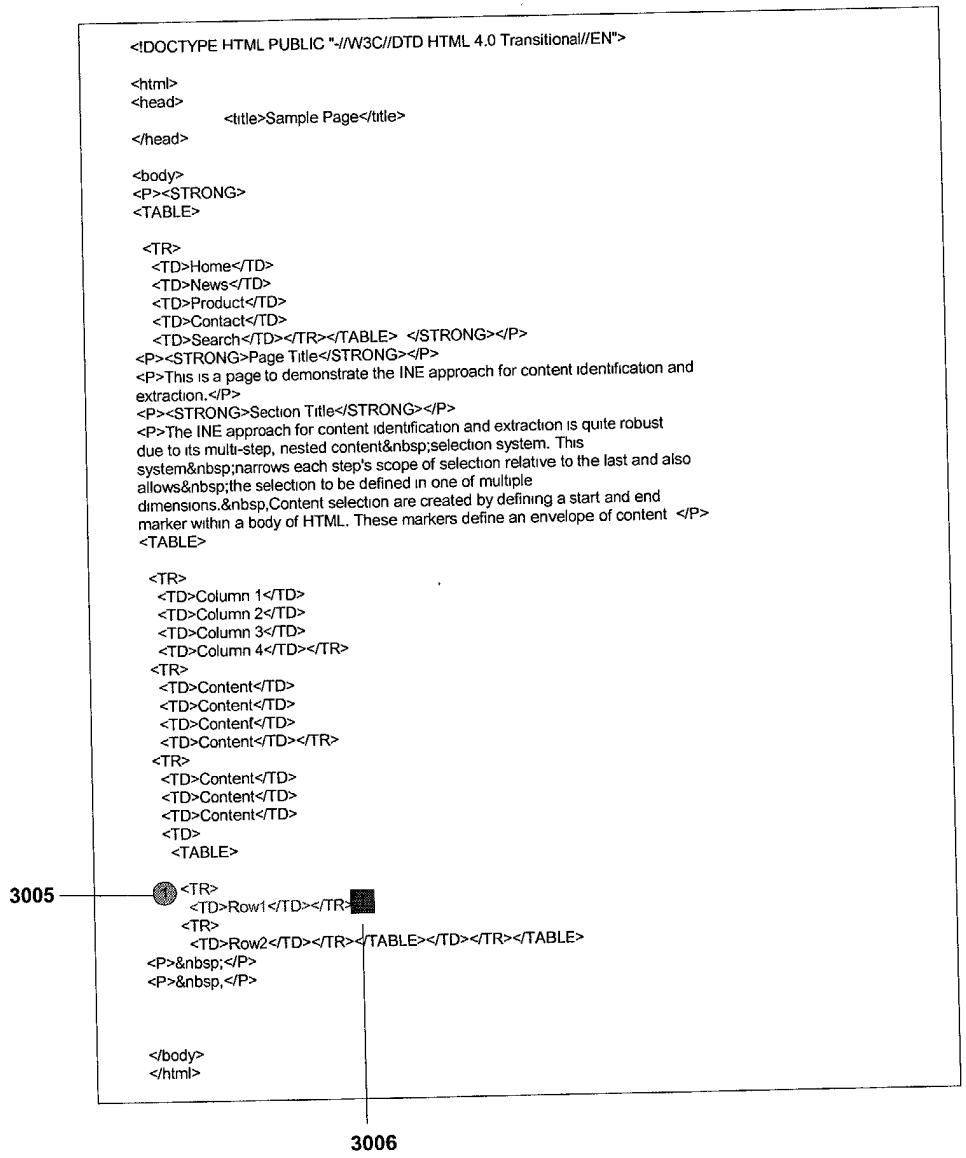

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<title>Sample Page</title>
</head>
<body>
<P><STRONG>
<TABLE>
<TR>
<TD>Home</TD>
<TD>News</TD>
<TD>Product</TD>
<TD>Contact</TD>
<TD>Search</TD></TR></TABLE> </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and extraction.</P>
<P>This is a new paragraph</P>
<HR>
<TABLE>
<TR>
<TD><STRONG> Section Title</STRONG></TD></TR>
</TABLE>
<P>The INE approach for content identification and extraction is quite robust due to its multi-step, nested content ,selection system This system ,narrows each step's scope of selection relative to the last and also allows ,the selection to be defined in one of multiple dimensions ,Content selection are created by defining a start and end marker within a body of HTML. These markers define an envelope of content. </P>
<TABLE>
<TR>
<TD>Column 1</TD>
<TD>Column 2</TD>
<TD>Column 3</TD>
<TD>Column 4</TD></TR>
<TR>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD></TR>
<TR>
<TD>Content</TD>
<TD>Content</TD>
<TD>Content</TD>
<TD>
<TABLE>
<TR>
<TD>Row1</TD></TR>
<TR>
<TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> </P>
<P> </P>
</body>
</html>
3007
3008
3009
3010
3011
3013
3014
3012

Figure 22




```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html>
<head>
        <title>Sample Page</title>
</head>

<body>
<P><STRONG>
<TABLE>

 <TR>
  <TD>Home</TD>
  <TD>News</TD>
  <TD>Product</TD>
  <TD>Contact</TD>
  <TD>Search</TD></TR></TABLE>  </STRONG></P>
<P><STRONG>Page Title</STRONG></P>
<P>This is a page to demonstrate the INE approach for content identification and
extraction.</P>
<P><STRONG>Section Title</STRONG></P>
<P>The INE approach for content identification and extraction is quite robust
due to its multi-step, nested content selection system. This
system narrows each step's scope of selection relative to the last and also
allows the selection to be defined in one of multiple
dimensions. ,Content selection are created by defining a start and end
marker within a body of HTML. These markers define an envelope of content  </P>
<TABLE>

 <TR>
  <TD>Column 1</TD>
  <TD>Column 2</TD>
  <TD>Column 3</TD>
  <TD>Column 4</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD></TR>
 <TR>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>Content</TD>
  <TD>
   <TABLE>

    <TR>
     <TD>Row1</TD></TR>
    <TR>
     <TD>Row2</TD></TR></TABLE></TD></TR></TABLE>
<P> </P>
<P> ,</P>



</body>
</html>
```

Figure 23

20220T0" 00E9500T

January 16, 2002¶

4000

Arthur Andersen Fires an Executive for Enron Orders¶

By RICHARD A. OPPEL Jr. with KURT EICHENWALD¶

WASHINGTON, Jan. 15 — Arthur Andersen fired its partner in charge of auditing the Enron Corporation today, saying he had ordered the destruction of thousands of documents and e-mail messages after learning that the Securities and Exchange Commission had begun an investigation of Enron's accounting. ¶

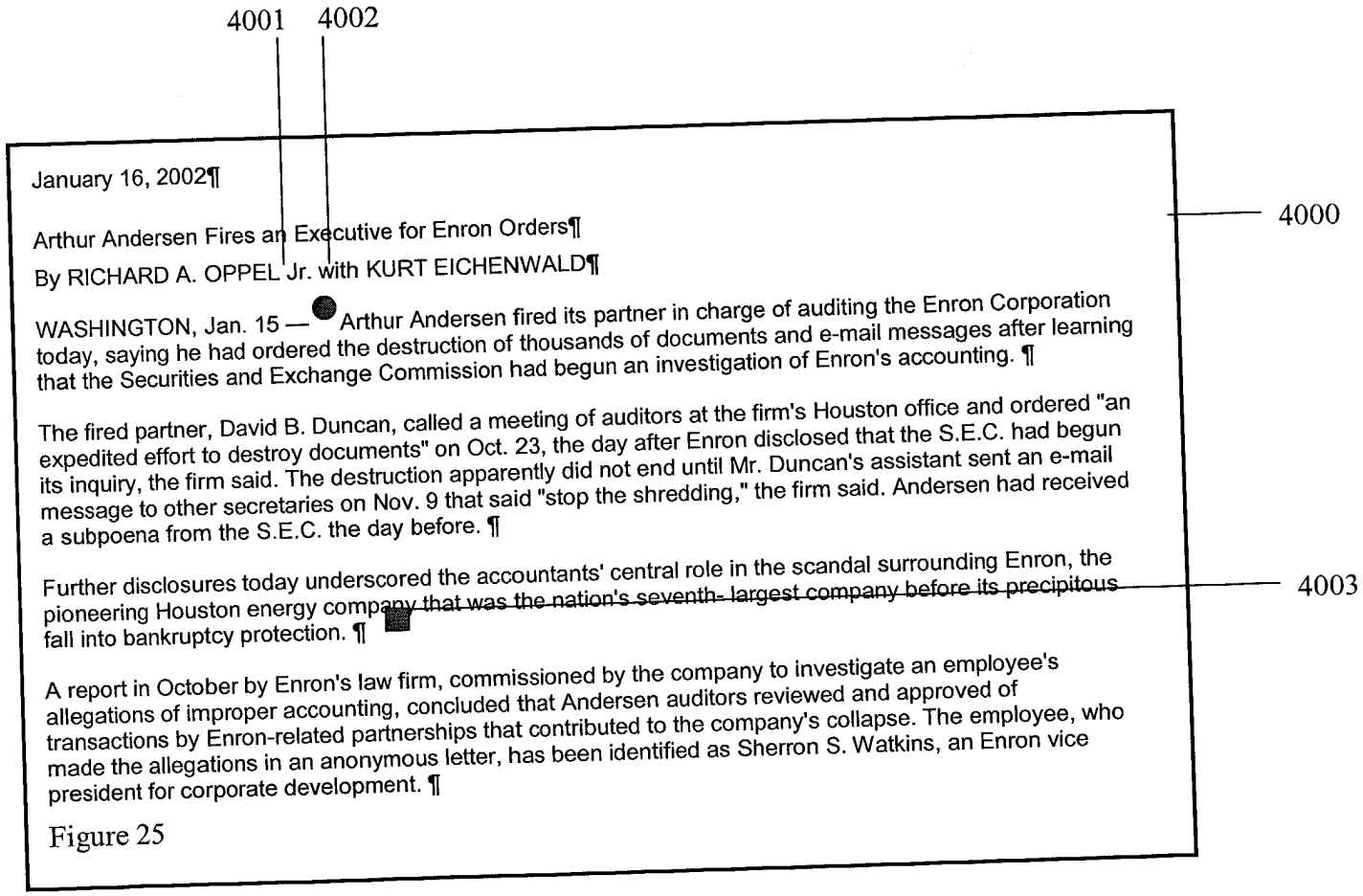
The fired partner, David B. Duncan, called a meeting of auditors at the firm's Houston office and ordered "an expedited effort to destroy documents" on Oct. 23, the day after Enron disclosed that the S.E.C. had begun its inquiry, the firm said. The destruction apparently did not end until Mr. Duncan's assistant sent an e-mail message to other secretaries on Nov. 9 that said "stop the shredding," the firm said. Andersen had received a subpoena from the S.E.C. the day before. ¶

Further disclosures today underscored the accountants' central role in the scandal surrounding Enron, the pioneering Houston energy company that was the nation's seventh- largest company before its precipitous fall into bankruptcy protection. ¶

A report in October by Enron's law firm, commissioned by the company to investigate an employee's allegations of improper accounting, concluded that Andersen auditors reviewed and approved of transactions by Enron-related partnerships that contributed to the company's collapse. The employee, who made the allegations in an anonymous letter, has been identified as Sherron S. Watkins, an Enron vice president for corporate development. ¶

Figure 24

20223TO "00E95007"



January 16, 2002¶

Arthur Andersen Fires an Executive for Enron Orders¶

By RICHARD A. OPPEL Jr. with KURT EICHENWALD¶

WASHINGTON, Jan. 15 — Arthur Andersen fired its partner in charge of auditing the Enron Corporation today, saying he had ordered the destruction of thousands of documents and e-mail messages after learning that the Securities and Exchange Commission had begun an investigation of Enron's accounting. ¶

The fired partner, David B. Duncan, called a meeting of auditors at the firm's Houston office and ordered "an expedited effort to destroy documents" on Oct. 23, the day after Enron disclosed that the S.E.C. had begun its inquiry, the firm said. The destruction apparently did not end until Mr. Duncan's assistant sent an e-mail message to other secretaries on Nov. 9 that said "stop the shredding," the firm said. Andersen had received a subpoena from the S.E.C. the day before. ¶

Further disclosures today underscored the accountants' central role in the scandal surrounding Enron, the pioneering Houston energy company that was the nation's seventh- largest company before its precipitous fall into bankruptcy protection. ¶

A report in October by Enron's law firm, commissioned by the company to investigate an employee's allegations of improper accounting, concluded that Andersen auditors reviewed and approved of transactions by Enron-related partnerships that contributed to the company's collapse. The employee, who made the allegations in an anonymous letter, has been identified as Sherron S. Watkins, an Enron vice president for corporate development. ¶

Figure 25